THE STRATEGIC NEWS SERVICE ©

GLOBAL REPORT ON TECHNOLOGY AND THE ECONOMY ™

SPECIAL LETTER

UNEXPECTED FLOWS AND INTERACTIONS IN THE DATA CENTER OF THE FUTURE

by David Binger

SNS SPECIAL LETTER: UNEXPECTED FLOWS AND INTERACTIONS IN THE DATA CENTER OF THE FUTURE

BY DAVID BINGER

In This Issue

Week of 07/09/2018 Vol. 23 Issue 22

Publisher's Note: If one were looking for a planetary dark side to the explosion of data centers and the related movement of corporate IT functions to the cloud, it would likely be the concomitant explosion in energy consumption, much of it ending up as waste heat.

Making even small-percentage dents in this problem provides huge financial – and ecological – leverage for customers, and for reducing global-warming issues.

Unfortunately, most of the technology addressing this issue seems vaguely like a rehash of automotive history, with air- and water-cooled systems at the forefront.

For that reason, I'm delighted to be able to share what appears to me to be a major step forward, using new science and new technology, to achieve markedly improved results on this front. Anyone owning or using data centers will want to read further, and then consider how to make use of this new invention. – *mra*.

SPECIAL LETTER:

Unexpected Flows and Interactions in the Data Center of the Future

by David Binger

Data centers are essential to the information economy. As the network extends sensors, computers, and displays outward to reach more places and more people around the planet, the demand for data centers will continue to grow. James Hamilton, an Amazon Web Services VP and Distinguished Engineer, estimates that global demand will not be satisfied until the number of medium to large data centers grows by three orders of magnitude, from hundreds to tens of thousands.[1] Whether or not that much demand is realized, it is clear that data-center construction will probably continue at a brisk pace for decades to come.

This huge growth in data centers will presumably deliver great benefits to people around the world, but the build-out will not be free of costs. This letter examines the problems of data centers with particular attention to the energy and mass that flows through them. In current data centers, these flows are managed by complex systems built on old technology whose advancement is widely understood to be completed. Fortunately, innovation is still alive. This letter presents a fundamental advancement in cooling technology that will simplify energy flows in data centers and significantly reduce the substantial financial and environmental costs of the cloud.

Two Big Problems: Power and Water

Unfortunately, data centers use very large amounts of power and water. An industry survey in 2013 put total global data-center power at 38GW, with 7% annual growth, which suggests a current total of about 50GW.[2] In the US, data centers reportedly consume around 2% of all generated electricity.[3]

The obvious problem with using so much power is that generating electricity creates pollution and dumps large amounts of carbon dioxide into the atmosphere. Data-center companies often invest in renewable-energy sources, but they still depend on continuing operation of generation plants fueled by coal and gas to meet

---

[1] James Hamilton, "How Many Data Centers Needed World-Wide," https://perspectives.mvdirona.com/2017/04/how-many-data-centers-needed-world-wide/

[2] DatacenterDynamics, "DCD industry census 2013: Data center power," http://www.datacenterdynamics.com/critical-environment/dcd-industry-census-2013-data-center-power/84829.article

[3] US Department of Energy, "Data Centers and Servers," https://www.energy.gov/eere/buildings/data-centers-and-servers

their needs when the sun goes down and the wind stops. Atonement can be purchased, but carbon footprints are not easy to erase. Pollution is a risk for everyone on Earth, but for high-profile cloud providers, pollution is also a financial risk, because their company valuations depend in part on maintaining a positive public image.

Water consumption is another problem for data centers. Most people don't know it, but power generation consumes vast quantities of fresh water. The US Department of Energy estimates that 39% of fresh-water consumption in the US is attributable to power generation plants. For comparison, that amount of water consumption is similar to the 41% that is used for irrigation and livestock.[4]

The water problem also extends to individual data centers, where water is heavily used for evaporative cooling. A 15MW data center, for example, may consume 130 million gallons of water each year.[5] The total "water footprint" of a data center is a problem of growing concern as populations grow faster than water supplies. This water usage creates a direct business risk to data centers. When water supplies run low, local communities will allocate water for uses that are most important to the local community. Data centers will probably be the first customers to have their taps turned off. Naturally, data-center owners would like to avoid the risks that go along with dependence on constant access to water.

A Critical Constraint: Heat Removal

As electricity flows through a computer chip, the electrical energy is converted to heat energy. The chip can't continue to operate unless an adequate path is provided for the heat to escape. Data centers house hundreds of thousands of processors, and they can't operate unless all of the heat from all of those chips has a way to escape out of the building and into the outside environment.

The heat energy flows out of the chip by conduction through the solid material, until it eventually gets to a surface that is exposed to a coolant – either air, water, or a liquid refrigerant. The heat energy moves into the coolant when molecules of the coolant interact with the solid surface. The warmed coolant moves away and takes the heat energy with it. The movement of the warmed coolant is generally driven by a fan or a pump, or else by the action of boiling or evaporation.

The coolant flow can circulate in an open loop or a closed loop. In an open loop, the coolant is drawn from and returned to the outside environment at a higher temperature. If the coolant circulates in a closed loop, the heat energy it absorbs at one point in the loop must be transferred out at the other side of the loop.

---

[4] National Renewable Energy Laboratory, "Consumptive Water Use for US Power Production," https://www.nrel.gov/docs/fy04osti/33905.pdf

[5] The Wall Street Journal, "Data Centers and Hidden Water Use," https://www.wsj.com/articles/SB10007111583511843695404581067903126039290

For example, the air that removes heat from the fins of a processor's heat sink typically circulates out of the rack and eventually to an air handler, where it dumps heat to the metal surfaces of a heat exchanger. From those heat-exchanger pipes, the heat is transferred to chilled water flowing in another loop. At the other side of the water loop, the heat from the water is transferred to another loop, this time with a refrigerant as the coolant. At the other side of the refrigerant loop, heat is transferred out of the pipes and into the water of a cooling tower.

In the last step, water moves in an open loop as it evaporates out into the outside environment, and a fan also blows air through the cooling tower in an open loop. The resulting system schematic is shown in Figure 1.

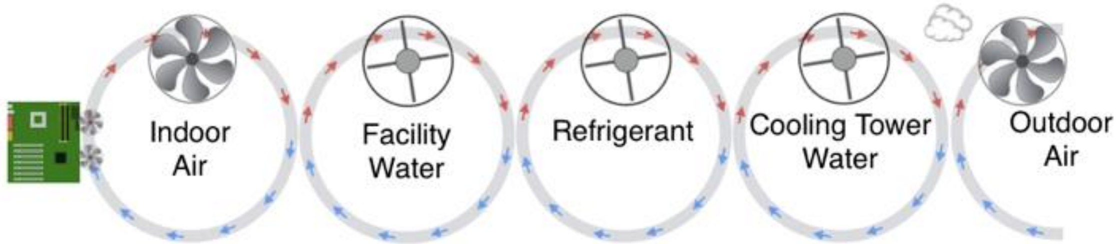

Figure 1: The typical system of cooling loops which removes heat from a server processor

These heat removal systems are complex. They require extensive plumbing systems, large industrial machines, and constant maintenance. The number of components involved is very high, and systems must be built with redundancy because component failures are expected.

Not all data-center heat removal systems are the same. Some are less complex than the one shown in Figure 1. Some are even more complex. Some utilize boiling or other phase-change mechanisms for transporting coolant. In almost all current data centers, though, you will find high numbers of moving parts, closed loops of circulating coolant, and reliance on the basic transfer of heat between a solid and a coolant.

The Closed-Loop Penalty

Closed-loop cooling methods have a common property: they transform a problem of moving heat from a smaller mass at a high temperature into the more difficult problem of removing heat from a larger mass of fluid at a lower temperature. Because heat transfer is driven by temperature differences, cooling gets harder when temperature differences get smaller.

As more closed loops are added to the system, the heat exchangers get progressively larger, more complex, and more expensive. That's the closed-loop penalty.

Troubles with Liquid Coolants

Liquid coolants have been available and sometimes used for cooling computers for as long as computers have been built. Throughout the decades, liquid cooling has been touted as the method of the future. Advocates of liquid cooling have claimed that it is the inevitable choice, because liquids are, they say, thousands of times better than air for cooling. Liquids do have a higher specific heat capacity and they can be a thousand times as dense, but these property differences have not boosted power densities by a thousand times.

Electronics cooled by liquids can, in fact, have higher power density than those cooled by conventional air cooling methods, but the improvements have not proven sufficient to drive wide adoption. Advocates of liquid cooling are understandably frustrated by the market's judgment, which they blame on deep-seated "hydrophobia."

Fear of liquids in the data center does have some justification. Water and high-density power are a dangerous mixture and can cause fires. Other liquid coolants may introduce leak risks, environmental hazards, risks of chemical exposure, and even the risk of asphyxiating people working in the data center. While it may be possible to reduce all of these risks to acceptable levels, data-center operators have always preferred to keep liquids away from their electronics.

Due to its higher specific-heat capacity, water is harder to heat up than air, and that property is helpful for removing heat from a processor. For the same reason, however, water is also harder to cool down than air, so the advantage at the chip is a disadvantage on the other side of the loop, where the heat must move out. This property of liquids means that, when they are used as coolants, the closed-loop penalty is higher than it is when the coolant is air.

Free Cooling

Some modern data centers have already worked around the closed-loop penalty. Instead of circulating air back and forth between servers and air conditioners, they open the building and bring in a constant supply of fresh air. The fresh air comes in one side of the building, and the hot air that comes out of server racks is moved by fans back to the outside environment. This open-loop "free cooling" approach isn't really free, but it may help to reduce costs, because the need for mechanical cooling systems is reduced.

In Lockport, New York, Yahoo built several data centers using a "chicken coop" design that brings in outside air through outer walls and blows the heated air out through the elevated central section of a barnlike structure.[6] On hot days, the air is

---

[6] Data Center Knowledge, "Inside the Yahoo Computing Coop," http://www.datacenterknowledge.com/archives/2010/09/20/inside-the-yahoo-computing-coop

cooled down by mechanical cooling units as it enters the building. In the Yahoo data centers, the air is moved through racks of servers by server fans.

The arguments against free cooling are that it is relatively complicated to manage, it won't work well in locations with hot weather, and it is unable to support equipment with high power density.

The Shocking Cost of Server Fans

Data centers of today typically use a very large number of small plastic fans to push air through server enclosures. A common example is a 40mm fan that draws 18W. A typical server might include six of these fans, and the cost to purchase all six will be around $100. Over the 10-year life of the data center, new servers will need to be purchased about three times, so the total purchase price for fans will be $300. Those fans must operate in a data center where the power and cooling infrastructure is very expensive. The low end of the cost range for this infrastructure is $6.44 per watt, so the cost for the 108W of all six fans is about $700. Over 10 years, operating the six fans will add over $1,000 to utility bills. The sum of purchase, infrastructure, and utility costs for a single server over 10 years is over $2,000. For a 15MW data center, the total cost for using servers with server fans is $30M.

Better Metrics Will Drive Better Decisions

For over a decade now, the data-center industry has used a metric of efficiency called PUE, or power usage effectiveness. This is the formula for PUE:

$$\text{PUE} = \frac{\textit{Total Facility Energy}}{\textit{IT Equipment Energy}} = 1 + \frac{\textit{Non IT Facility Energy}}{\textit{IT Equipment Energy}}.$$

If a facility uses 100% of its energy for IT equipment, the PUE is a perfect 1.0. In real data centers, however, energy is used for converting power to lower voltages, for providing uninterruptible power supplies (batteries), for lighting, and for powering the mechanical cooling systems of the building. The biggest part, by far, of this energy overhead is the part that provides cooling, so we in the cooling equipment business sometimes use a very similar metric, called cooling PUE, or mechanical PUE (mPUE), defined as:

$$\text{mPUE} = 1 + \frac{\textit{Non IT Energy for Facility Cooling}}{\textit{IT Equipment Energy}}.$$

In marketing materials for data-center cooling equipment, it is common for vendors to use the term PUE when they are actually referring to mPUE.

When the outside air is hot, facility cooling systems need to work harder, and the PUE values are higher. When the outside air is cooler, PUE values fall. For a particular data-center location, PUE is usually reported as an average of hourly values over a one-year sampling period.

When the PUE was first introduced in the industry, data centers were not very efficient. PUE values of 2.0 or more were common, which meant that a data center would use as much energy for cooling systems as it used for powering IT equipment.

Perhaps because companies started reporting PUE values, data-center designs and cooling equipment began to change, and PUE values improved. Now, a conventional data center may typically report a PUE of 1.4.

The large "hyperscale" data centers of Amazon Web Services, Microsoft, Google, Facebook, and Apple have improved PUEs even more than ordinary data centers by taking advantage of scale, uniformity, and total control of everything from IT equipment to facility to site selection. Their reported PUE values have been as low as 1.07.

The establishment of the common metric was followed by a period of improvement, so that we now have PUE values that are much better than before, and facility cooling systems are not using as much energy as they once did. That's pretty good news, but there are a couple of unintended consequences that are not good.

The first problem is the above-mentioned issue of the consumption of fresh water. Data centers have achieved their efficiency gains by using less power-hungry compression cooling and more evaporative cooling systems, which consume lots of water on the data-center site. The defense of this choice, from a water consumption point of view, is that the alternative of using more power-hungry compressors would cause even more fresh-water consumption at the power generation plant. Nevertheless, data centers that are dependent on a continuous supply of cooling water are exposed to an extra risk of interrupted operation.

The second problem with PUE metrics is in the denominator of the equation: *IT Equipment Energy*. Power to the IT equipment is measured at the point that it enters the racks. That means it includes power conversions that occur in the racks as well as power used to operate server fans. The metric counts server fans and in-rack power conversions as benefits, but they should be counted as costs. If you install more-efficient power supplies and fans in your rack, your PUE score gets worse. The flawed metric motivates some bad choices.

Everyone in the industry knows that the PUE has these and other flaws, but PUE is the incumbent efficiency metric, and change does not come easily. We should, however, look for better metrics.

Power density is another metric that's frequently used in marketing IT equipment and data centers themselves. With more power in a rack, fewer metal rack structures and less floor space are needed. This saves money.

For data centers, though, the business objective is to produce net revenue by providing computer operations at a low cost. The metric that seems to matter most is the total of all direct and indirect costs for a given quantity of computer operations.

Flows and Interactions

SNS members are probably already aware of Mark Anderson's idea that it's useful to think about things in the world in terms of flows and interactions. This way of thinking can help us understand the important processes in a data center.

In a CPU, each transistor is managing a flow of electrons. The electrons interact with one another and also with the molecules in the chip. These interactions cause the molecules in the chip to move around faster – so the chip heats up, and that's where the heat removal problem begins.

You can increase the performance of the chip by increasing the clock frequency. Each state change requires a bigger flow of electrons and higher voltage. This bigger, more energetic flow of electrons creates more and stronger interactions with molecules. Consequently, even more heat needs a way out.

We can also look at the flow of thermal energy, transmitted by interactions between molecules of conducting solids and those of coolants until it escapes the confines of the data center.

At a larger scale, we can look at the mass flows of air, water, and refrigerant as they move through their respective cooling loops, with warming interactions and cooling interactions allowing the thermal energy to move through the cooling system.

The Critical Interactions of the Cooling System

The critical interactions of a cooling system are those between a hot solid surface and molecules of the coolant, be it air, water, or refrigerant. These interactions transfer heat energy from the solid to the coolant.

Historically, all cooling systems have had only one possible method to support higher power demands. The only available option has been to increase the number of these critical interactions at the boundary between the solid surface and the coolant. For example, more powerful chips need bigger heat sinks so there is a larger surface area for air molecules to contact. Alternatively, the coolant is changed to a high-density substance such as water, so that more coolant molecules come into contact with the hotter solid surface.

To date, more power has always required a larger quantity of critical interactions. There has been no progress at all in improving the quality of these interactions, and no such progress has been anticipated.

The Surprising Future of Data Centers

Scott Davis, CEO of the company Forced Physics (*Disclaimer: I am the company's CTO*), has discovered a way to create a physical phenomenon that changes the interactions between gas molecules and a solid surface. His invention is the first to exploit this physical phenomenon to provide cooling. The invention enhances the transfer of heat from a solid into a gas. This enables more heat to be transferred with smaller flows, because the interactions are more productive.

Davis formed Forced Physics to secure the intellectual property and commercialize his invention; the company is now producing products that cool data-center electronics. The new approach eliminates all closed loops from the data-center cooling system. In contrast to the complexity shown in Figure 1, the flow in the new system has the following structure:

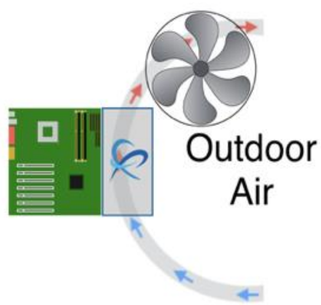

Figure 2: The Forced Physics solution for cooling electronics

The new system uses a very compact and effective heat exchanger to transfer heat, conducted from electronic components, directly into a single open loop of unchilled (but filtered) air from outside. The 150°F exhaust is blown directly back into the outside environment. Owners may decide to use the hot exhaust for some additional benefit, such as generating additional electrical power.

This approach dramatically simplifies the data center. It eliminates the costs and risks of server fans, chiller plants, plumbing, and water towers. It reduces the number of moving parts in the data center by three orders of magnitude. It reduces total redundant power capacity required in the data center. As a result, the costs of the facility go down significantly.

In this system, the only power required for cooling is what is needed to operate the high-efficiency industrial blowers that move the air through the heat exchangers. The power for the blowers will average only about 2% of the server heat. An average mechanical PUE of 1.02 will be obtained, even in a hot climate like that of Phoenix, Arizona. Total electrical energy usage is reduced significantly, and the carbon footprint of the data center shrinks proportionally.

Cost savings even extend to the server equipment, because the new system supports printed circuit boards with a very high-power density. Substantial savings come from increasing performance and power from existing chips and from aggregation of electronic components.

By using this new cooling technology, data centers of the future will be simpler, cleaner, and cheaper. They won't use any water for cooling, and they won't use any refrigerants. The only coolant will be air.

Server fans will disappear from data centers. Servers and other IT equipment will have very high-power density, and the cost of providing computer operations will be lower than anyone expected.

About David Binger

David Binger, PhD, joined Forced Physics in 2013 and serves as the company's chief technology officer. David and Scott Davis, inventor and CEO, constructed and tested hundreds of prototype cooling devices with the goal of finding a scalable and manufacturable design that delivers outstanding heat transfer performance. In 2015, the team produced a laboratory-scale prototype that demonstrated high-density cooling.

Before joining Forced Physics, David worked at the MEMS Exchange, a DARPA project hosted at the Corporation for National Research Initiatives (CNRI) in Reston, Virginia. As a senior software developer, he led a team of developers that created a custom ERP system to manage the unique and highly complex workflows of the world's first distributed MEMS fabrication network. Earlier, at the University of Illinois at Urbana-Champaign, David earned a PhD in Computer Science in the area of electronic design automation algorithms.

I would like to thank David for taking the time to share the details of both the problem and of this new solution. SNS members attending FiRe can meet David and the rest of the Forced Physics team there, where they will be one of our FiReStarter companies for FiRe 2018.

And, last but never least, our gratitude to Editor-in-Chief Sally Anderson, for putting all of these thoughts into perfect shape.

Your comments are always welcome.

Sincerely,

Mark R. Anderson

CEO
Strategic News Service LLC
P.O. Box 1969
Friday Harbor, WA 98250 USA

Tel.: 360-378-3431
Fax: 360-378-7041
Email: mark@stratnews.com

CLICK HERE TO SHARE THIS SNS ISSUE

To arrange for a speech or consultation by Mark Anderson on subjects in technology and economics, or to schedule *a strategic review* of your company, email mark@stratnews.com.

We also welcome your thoughts about topics you would like to suggest for future coverage in the SNS Global Report.

For inquiries about **Partnership or Sponsorship Opportunities** and/or SNS Events, please contact Sharon Anderson Morris, SNS Programs Director, at sam@stratnews.com or 435-649-3645.

INSIDE SNS

Please visit www.stratnews.com/insideSNS for:

- Photo galleries of FiRe and other SNS events
- FiRe videos

- SNS iNews®
- The SNS blog, "A Bright Fire"
- The SNS Media page
- SNS FiReFilms
- Subscription rates and permissions
- About SNS and About the Publisher

UPCOMING SNS EVENTS

Credit: Kris Krug (L) and David Morris

Register now for FiRe 2018
The 16th annual Future in Review conference
October 9-12

Returning to the beautiful
Stein Eriksen Lodge Deer Valley
Park City, Utah

www.futureinreview.com/register

WITH GREAT APPRECIATION TO:

Our Global Platinum and FiReFilms Partner:

Global Platinum Partner:

Global Gold and FiReFilms Partner:

Global Silver Partner:

and Focus Channel Partners:

... for their Partnership and Support of SNS events.

ADDITIONAL SUPPORTING ORGANIZATIONS

USTAR

FiReFellows Sponsor:

ZIONS BANK

and FiRe Academic Partner:

SCI
www.sci.utah.edu

Where's Mark?

• On October 9-12, Mark will be hosting the 16th annual Future in Review (FiRe) conference, at Stein Ericksen Lodge Deer Valley in Park City, Utah. Special guests will be joining us from the governments of Taiwan, Germany, Sweden, and Ukraine, with business leaders from Australia, the UK, and New Zealand.

To register now for this 200-person SNS event, go to:

www.futureinreview.com

ISSN 1093-8494